图书情报工作
第64卷 第5期 2020年3月

多源数据环境下科研热点识别方法研究

■ 裘惠麟[1] 邵波[1,2]
[1] 南京大学信息管理学院 南京 210046 [2]南京大学图书馆 南京 210046

摘 要：[目的/意义]在科学研究中，从不同来源的科技文献中识别挖掘科研热点对于开展科研工作具有指导意义。旨在通过本研究提出的模型方法，快速准确地识别蕴含在多源文本中的热点主题，为科研创新提供支撑服务。[方法/过程]提出一种基于LDA2vec模型的多源文本下科研热点识别的方法并针对科研热点识别构建模型，该方法融合LDA主题模型对隐含语义挖掘的优势和Word2Vec词向量模型对于上下文关系把握的优势。以机器学习领域的科技文献为例，利用模型困惑度和主题一致性两个指标对LDA2vec的在本领域应用的可行性和有效性进行验证，并与LDA的主题提取效果进行对比。[结果/结论]实验结果表明，提出的方法在面对多源数据情况下，进行科研热点识别挖掘是可行的，且在一定程度上有效果的提升，对利用单一数据源进行主题分析的不足进行补充，对多数据源融合的实践应用进行丰富。

关键词：主题模型 LDA2vec 科研热点 LDA Word2vec 多源数据融合
分类号：G251.2
DOI：10.13266/j.issn.0252-3116.2020.05.009

信息爆炸增长的态势随着技术和时代背景的发展愈演愈烈。在互联网上进行信息检索和收集时，除了有效信息，也会被大量无用无关的信息干扰。在科研工作中，面临的情况也是如此。对学科领域内已有的研究成果、期刊论文的阅读与研究，是科研工作者在短时间了解把握学科研究现状、形成对学科领域较为全面认知的一个主要的手段[1]。因此，要能够及时把握研究现状、跟进主要的研究热点与方向，热点识别与挖掘是一个有效且可行的办法。但目前大多数研究中对于科研热点的识别主要针对期刊论文，单一的数据源得出的分析结果必不能全面地反映学科领域的整体研究现状。论文与专利分别反映的是基础研究和技术创新成果的进展情况，虽然两者在文献结构及文字表达上存在差异性，属于异构文献，但在内容上可以实现有效整合形成新的技术信息，与单一文献源相比，在信息的全面性、科学结构划分的准确性上更有优势，对于准确定位领域的研究重点、热点和预测领域研究趋势都大有裨益[2]。将两者结合起来进行主题热点分析，对于理解科学和技术的相互影响和渗透关系、技术机会识别、潜在商业化机会发现等方面有着重要的意义。

如果面对多源数据，不同源的文本本身存在异构性、且大概率不会存在引用关系时，图情学科内传统的计量分析方法、基于关键词与主题词等的分析方法就不能有效地得出结果。为了尽量避免传统方法带来的较为宏观、粗糙的结果，解决停留在文献外部特征分析带来的全面、客观性不足、层次不够深入的问题，现在越来越多的研究中采用了文本挖掘的方法。基于对文本内容的挖掘，能够更加有效地对文献的内部特征做到客观、全面的识别与分析，能对研究的粒度与层次有一定的提升。面对识别分析多源文本的需求，基于主题模型的方法不仅能够更好、更方便地解决多源文本在异质、异构上的问题，更多的是关注科技文献的文本内容，获得综合性的分析判断结果，增加结果的置信度；也能更深层次地挖掘文本内在的隐含性知识，突破词频分析等外部特征统计，利用机器学习的便利性，在更短的时间内理解文献要表达的内容，并挖掘出更多、更丰富的语义信息与知识推理，这对科研工作乃至情报分析研究工作，都有着更高更优的全局性价值。因此，主题模型在科研热点识别领域的应用与优化是值得探索研究的。

作者简介：裘惠麟（ORCID：0000-0001-6516-2789），硕士研究生，E-mail：qiuhuilinqha@163.com；邵波（ORCID：0000-0002-6528-5196），副馆长，教授，博士。
收稿日期：2019-05-20 修回日期：2019-09-16 本文起止页码：78-88 本文责任编辑：徐健

1 相关研究

1.1 科研热点识别方法

图情领域的科研热点发现，依赖于对不同实体之间隐藏关系的发现与推理，通过关系发现隐形知识与前沿热点等，是对科学计量学、情报分析与研究等的重要拓展与实践，也一直是学科内重要的研究领域。科研工作者在进行研究热点的识别与分析时，主要使用的方法可以归纳为基于文献外部特征的方法和基于文献内部特征的方法。

基于文献外部特征的方法包括引文分析法和知识单元分析法。引文分析是以文档间引文的频率和模式为研究对象，通过引用模式——从一个文档到另一个文档的链接，以揭示文档的属性[3]。除直接引用外，共引分析和耦合分析目前在研究热点识别中的应用比较广泛，且在原本方法的基础上，有了诸多的发展，如共引中的文献共引、词的共引、主题共引、作者共引和类的共引等[4]，耦合中的文献耦合、作者耦合[5]、关键词耦合、期刊耦合[6]等。知识单元作为构成知识集合系统的最基本单位，在科学计量研究界，可以狭义地理解为不能再分解的词[7]。因此，笔者将基于词频统计的分析方法和主题词共现的分析方法归于知识单元分析法，这两类方法的主要特征就是基于文献的最基本单元(词汇)的外部特征，相比于引文分析，更能从微观层面揭示学科结构内的实体关系[8]。知识单元分析方法在研究热点发现领域应用非常广泛，但也有部分学者认识到利用被引频次、引文关系等的引文分析并不能够直接展现文献的研究内容，如祝清松和冷伏海认为基于引文内容分析的主题在揭示高被引文献的被引原因上效果更好，并且与论文的整体内容相符[9]。

基于文献内部特征的方法可以理解为基于文本内容挖掘热点的方法。从语义层面进行挖掘以识别文档集的主题与内涵，能够在一定程度上解决基于文本外部特征分析而出现无意义、偏离文本原意结果的问题。如杨超[10]通过抽取专利文本中的SAO结构构建主题模型，在识别专利文献主题时解决了主题语义不清、问题解决方案识别不对应的问题；阮光册[11]采用Doc2Vec方法对文本内容进行向量计算与相似度计算以生成热点选题论文集，在此基础上再利用主题模型和聚类算法进行主题识别与挖掘，在语义特征的识别上获得了更优的效果；赵一方[12]针对政策文本引入段落信息增益，改变了现有主题模型无法有效分配特定特征词对相似政策主题贡献度的问题，平衡了不同主题间的贡献差异。

整体来说，笔者认为目前在科研热点识别领域，相对成熟的方法体系已经初步形成，不仅包含本学科内如文献计量领域的自有方法，也包括从其他学科、以及新兴技术发展中引入的方法。但存在的主要问题包括3点：①语义理解不足问题，以传统科学计量为基础的方法大部分的核心思想都是对文献中的主题词进行统计(如词频、共现频率、共引次数等)，但对于文本和语义层面却没有深入研究(如同义词、近义词、不同词汇习惯等)；②研究数据源单一问题，以某一种数据源来识别研究前沿具有局限性，并不能全面代表所有科学研究前沿信息；③时滞性问题，学术论文从撰写、审稿到产出并形成引用关系的过程一般情况下十分漫长，所以这个过程本身就会使得论文数据在时间上存在滞后性。

1.2 主题模型研究现状

主题由一个核心事件或活动以及所有与之直接相关的事件和活动组成[13]。利用主题模型可以对文献进行内容分析、提取主题以获得领域内的热点知识和发展趋势。

从文献调研的结果来看，由于LDA模型本身是应用范围最为广泛又较为成功的模型，且其对于大规模文档集隐含语义的识别效果较好，而科研热点识别的核心就是从大规模的学科领域科技文献中挖掘、推理隐含知识，所以应用于科研热点识别的主题模型主要以LDA模型为基础[14]。当然，针对LDA模型实际应用于科研热点及主题识别时的一些问题，国内的诸多学者也进行了优化或将LDA与其他模型相结合以达到研究目的与效果：除了LDA模型和pLDA模型之外，还包括了将LDA与本体[15]、SNA社会网络分析法[16-17]、引文分析法[18]、共词法[19-20]、标签[21]、聚类算法[22]及相关特殊指标相结合[23-26]的优化方法。

所以从当前国内研究情况来看，整体来说将主题模型应用于科研热点识别的方向比较单一，大多数是在LDA模型的基础上进行相关的改进，对于一些新模型、新方法的可行性与有效性的探索却比较少。反观如舆情热点识别、微博热点识别等其他领域，探索的方法就较为多样化、丰富化。而很多的新方法在对文本语义的理解和挖掘上或许有着更好的效率和更优的效果，所以，我们进行科研热点识别与挖掘时，这些都值得探析。

2 基于 LDA2vec 的科研热点识别方法

2.1 模型基础

2.1.1 LDA 主题模型

主题模型是一种非监督的机器学习方法,它不同于图情学科内传统的基于文献外部特征的方法,传统的方法只关注文献之间的表层关系或词语频次等,而主题模型可以将词汇与文档之间深层次的语义关系抽取出来,也就是我们所说的"潜在主题信息",有效地提取大规模文档集和语料库中的隐含主题,目前已在文本情感分类、信息抽取等领域已经广泛应用。这为深入的进行文本分析、科研主题挖掘提供了很好的契机,有着广阔的应用前景和现实意义。从 1998 年最早的主题模型 LSI 提出以来[27],在此基础上至今有了很多优化的模型算法,通过对大量文档、句子、单词的计算与学习,能够对文档集合中隐含的语义结构进行探索。

主题模型的核心是为了文本降维,文本降维技术由 TF-IDF 矩阵、一元混合模型、pLSA 模型等发展到最经典的 LDA 模型,可以将其理解为对 pLSA 模型进行贝叶斯化,即 LDA 是由单词、主题和文档组成的一个三层贝叶斯网络模型。核心思想就是:每个文档可以被视为各种主题的混合,其中每个文档被认为具有通过 LDA 分配给它的一组主题。LDA 通过计算 P(单词|主题)和 P(主题|文档)来获得单词的分群。其中最关键的两个步骤是:①该词在所有文档的范围内归属于哪个主题;②该词所在的文档归属于哪个主题。总体来说,笔者认为 LDA 有两大好处:①能够处理多义词或者同一个词的不同语境。因为 LDA 进行主题划分时,又考虑到整个文档的主题倾向。②可以对每个主题,都找出一些词来描述它。这对于更全面、深刻地理解某一主题的含义有更好的指导作用,这在科学研究中也是大有裨益的。但 LDA 最大的劣势在于其是一种典型的词袋模型,它认为一篇文档是由一组词组成的集合,没有考虑到词与词之间的顺序和先后关系。

2.1.2 Word2vec 词向量模型

尽管在 LDA 中能大致对应于主题,但对于词向量通常不是这种情况。在主题模型为每个词语分配到的是一个和上下文语境、语义无关的向量,但是要深度理解文本的语义与内容,上下文语境却是需要着重考量的。LDA 模型未能将词与词之间的关系纳入考量与计算,而词向量模型的一大特点正是对词汇之间的关系进行描述。词向量和词的内容无关,而是和语义相关,更加关注上下文逻辑。

Word2vec 主要有两个模型:一是在词袋结构(CBOW)中,基于一组上下文词来预测枢轴词;二是在 Skip-gram 架构中,枢轴词用于预测周围的上下文单词图描绘了两种不同的 Word2vec 架构。也就是说,CBOW 中输入的是词 w_i 周围 n 个词语的向量之和,输出词 w_i 本身的向量;Skip-gram 中输入词 w_i 本身的向量,输出词 w_i 周围 n 个词语的向量。

结合模型本身来看,LDA 模型的基础是隐含主题,Word2vec 模型的基础是上下文。即 LDA 关注的核心是文档和词的共现,而 Word2vec 关注的核心是上下文和词的共现。两者对于语义分析来说是优劣互补的,也是本研究模型构建的基础。

2.2 模型构建

C. E. Moody 等[28]提出的 LDA2vec 模型是一种与 Dirichlet 分布的潜在文档级主题向量混合共同学习密集单词向量的模型,同时吸取了 LDA 模型对于主题把握的优势和 Word2vec 模型对于词语之间关系把握的优势,将两者融合,在 Word2vec 的 skip-gram 模型基础上建模,由本来的输入某个词语以预测上下文词语转变为使用上下文向量来预测上下文词语。即可以理解为在原始的 Skip-gram 方法中,训练模型以基于枢轴词来预测上下文词。在 LDA2vec 中,添加了枢轴词向量和文档向量以获得上下文向量,然后使用该上下文向量来预测上下文单词。具体来说,扩展了 Skip-gram 模型,融合主题和文档向量,并结合了词嵌入和主题模型的想法。受 Latent Dirichlet Allocation(LDA)的启发,将模型扩展为同时学习词、文档和主题向量。

所以笔者参考 C. E. Moody 等提出的 LDA2vec 模型,希望能通过更多的数据和更多的特征来对周边词汇进行更高效的预测,以更有效地提取隐含在文献内部的主题。笔者基于两者的混合模型 LDA2vec,借鉴其对 LDA 模型全局性和 Word2vec 模型局部关系进行整合利用的思路,探讨将稀疏文档表示与密集词和主题向量混合的热点主题识别方法,构建了如图 1 所示的模型。

在原始 Skip-gram 模型中,如果枢轴词是"机器学习",则可能的上下文词可能是"计算机""人工智能""算法"。如果没有任何全局性(文档相关)的信息,这些预测结果是具有一定的合理性的。但通过在 LDA2vec 模型中提供附加的上下文向量,也许可以更好地对上下文词语进行预测。

C. E. Moody 等对 LDA2vec 模型的实现算法对设

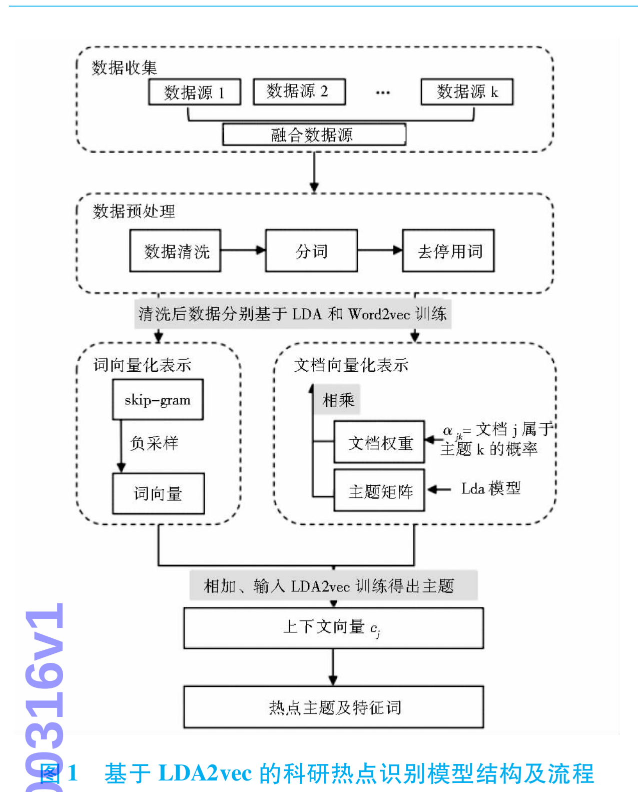


图 1　基于 LDA2vec 的科研热点识别模型结构及流程

备 GPU 要求过高，适用于超大规模数据，效率较低。从 github 上基于其模型的实验来看，实验结果与传统 LDA 相比差距并不明显。所以在本研究的实验中，考虑到热点识别的需要以及原始数据的规模并不大，笔者对模型实现进行了一些改进，先利用成熟的 Word2vec 模型和 LDA 模型对语料库进行训练，然后将其结果作为输入，利用 LDA2vec 模型中核心算法进行迭代计算，以期得到更优结果的同时，提高效率。

本模型通过处理文档并将文档向量分解为文档权重向量和主题矩阵。文档权重向量表示不同主题的百分比，而主题矩阵由不同的主题向量组成。因此，通过组合文档中出现的不同主题向量来构造上下文向量。即首先基于 Skip-gram 模型，提取在扫描语料库的移动窗口中出现的枢轴和目标词对。对于每个词对，枢轴词用于预测附近的目标词。其次对于语料库中的每个文档随机初始化潜在向量。文档权重是 softmax 变换的权重以产生文档比例。结果是一个比例向量，总和为 100%，表示单个文档的主题比例。例如，一篇文档可能包含了 3 个主题：主题 0 为 41%，主题 1 为 26%，主题 2 为 34%。

每个主题都有一个分布式表示，与单词向量位于同一空间。虽然每个主题在字面上并不是语料库中存在的标记，但它与其他标记类似。每个文档向量是主题向量的加权和。因此，这种分析可以产生可解释的主题，帮助人们直接理解文档的主要内容，不再需要详细阅读。

2.3　模型解析

本研究构建的整个模型中核心算法部分主要包括两部分的计算与训练：一部分用于训练得到某篇文章不同主题所占比重的信息；另一部分基于 Skip-gram 的方法，在枢轴词和目标词确定的情况下，学习上下文的向量表示。

2.3.1　词向量表示

词向量学习表面上包括两部分，首先根据 Skip-gram 得到词向量表示，之后引入上下文向量，采用 Skip-gram 负采样的思想学习目标词的词向量表示。但第二部分的词向量在本模型中是不改变的，实际上是希望借鉴词向量的训练方法最小化（枢轴词＋文档，目标词）对与（枢轴词＋文档，随机词）的损失函数从而学习内容的向量表示。

同时与之前的 Word2vec 中的方法一致，负采样时根据词频决定采样的可能性大小。某个单词采样的可能性大小如下，参数取 3/4：

$$\mathrm{len}(w)=\frac{[counter(w)]^{3/4}}{\sum_{u\in D}[counter(u)]^{3/4}}\qquad \text{公式(1)}$$

与 Word2vec 模型一样，当输入词和目标词对（j，i）在跨越语料库的移动窗口中共同出现时被提取。对于每个（输入词－目标词）词对，输入词用于预测其附近的目标词。每个单词用固定长度密集的分布式表示向量表示，但与 Word2vec 模型的不同在于，在输入和目标表示中使用相同的单词向量。绘制词语的分布是 u^{β}，其中 u 表示由总体语料库大小归一化的整体词频。除非另有说明，否则采样功率 β 设置为 3/4，负采样数固定为 n＝15。与 unigram 分布相比，这样的选择更强调了为负样本选择不常用的单词。与优化 softmax 交叉熵相反，负面采样通过从语料库中每个边缘的边缘流行度中抽取负样本来研究以上下文为条件的学习单词向量。

2.3.2　文档向量表示

这部分工作的意义在于，通过得到对应文档的文档向量表示之后，和相应的词向量相加，作为上下文向量的初始值。

某个单词 j 对应上下文的初始值设定如下：

$$\vec{c_j}=\vec{w_j}+\vec{d_i}\qquad \text{公式(2)}$$

其中，$\vec{w_j}$表示词语 j 的词向量，由前述步骤获得；$\vec{d_j}$表示对于词而言，所有词－上下文对的向量表示。其具体公式表示如下：

$$\vec{d_j}=a_{j0}\cdot\vec{t_0}+a_{j1}\cdot\vec{t_1}+\cdots\qquad \text{公式(3)}$$

$\vec{t_0}$,$\vec{t_1}$等表示对应主题的向量表示。基于 LDA 模型得到主题矩阵后,经过矩阵分解方法得到的与词向量长度一致的结果。a_{jk}表示对于文档 j 而言,属于主题 k 的概率,取值在 0 到 1 之间。

需要指出的是,主题向量表示对于所有文档是通用的,但不同文档里面具体的主题分布就是通过 a_{jk} 来决定的。为了保证 a_{jk}的可解释性,这里采用了 softmax 的方式保证其和为 1 且非负。同时在得到$\vec{t_1}$之后,基于词向量与该主题的相似程度可以得到相关主题词汇。

a_{jk}的具体计算与 L^d 密切相关,具体计算公式如下:

$$L^d = \lambda \sum_{jk} (\alpha - 1) \log p_{jk} \qquad \text{公式(4)}$$

当 $\alpha < 1$ 时,主题分布倾向于稀疏。反之,主题分布会更加集中。为了增强模型的可解释性,这里取 $\alpha = n^{-1}$,n 表示 topic 的数目。同时,当 $\lambda = 200$ 时,模型的表现效果较好。

3 实验结果及评估——以机器学习领域研究为例

3.1 数据来源与预处理

首先,选择一个学科发展相对成熟、边界比较清晰的学科为分析对象,笔者选择机器学习领域的研究成果为数据对象。于 2019 年 2 月 1 日在 CNKI 期刊数据库和专利数据库中分别检索发表的全部中文文献。根据本次实验的要求,查找机器学习学科领域的相关科技文献(包括学术论文和专利文献)。设定检索表达式为 SU =‘机器学习’,并限定文献出版时间与专利公开日均为 2004 年至 2019 年的 15 年期间,于 2019 年 2 月 1 日的检索结果为:共 5 869 篇期刊文献和 3 865 篇专利文献。对原始数据进行筛选,剔除与学术研究无关的内容和重复项后,汇总的数据共 8 928 条,其中包括 5 063 条期刊论文数据和 3 865 条专利文献数据。在 CNKI 数据库中,对于每篇期刊文献和专利文献都有题名和摘要的标引,这对于我们做多源文本的融合提供了支持。将期刊论文数据和专利文献数据按照题名项和摘要项汇聚在一起,形成初步原始数据集。

上述步骤后得到的这些内容是直接从数据库抓取的、未经处理的原始数据,需要通过分词和去除停用词,将原始的数据处理成便于后续模型输入、可供计算机识别理解的内容。笔者将采用 jieba 作为分词及去停用词的工具。jieba 分词工具可以除了进行分词以外,也支持对停用词的过滤去除,本次实验利用 jieba 将文档中没有具体实际含义的词以及标点符号等作为停用词进行过滤处理。文本预处理是一个需要重复进行的过程,对分词自定义字典进行扩充,进行特征选择直到处理结果能满足模型输入的要求。

3.2 基于 LDA2Vec 的主题提取

LDA2vec 模型的提出者 C. E. Moody 在 github 上开源了模型的核心库,但是从基于其模型的实验来看,首先是对 GPU 要求过高,其次是实验结果与传统 LDA 相比差距并不明显。所以在本研究的实验中,对模型实现进行了一些改进,先利用成熟的 Word2vec 模型和 LDA 模型对语料库进行训练,然后利用 LDA2vec 模型中核心算法进行迭代计算,以期得到更优结果的同时,提高效率。

3.2.1 词向量表示

笔者将前述预处理步骤后获得的论文和专利文本作为语料库,利用 Word2vec 来生成论文与专利文本融合文档集中词语的词向量,作为后续模型的输入。

Python 的 Gensim 工具包对 Word2vec 模型进行了封装,本实验基于 Python 的 gensim 包中 gensim. models. word2vec 类实现对于 Word2vec 中 Skip-gram 模型词向量的训练。根据本研究需求,对于 Word2vec 模型中相关重要参数设置如表 1 所示:

表 1 word2vec 模型参数设置

参数	值	原因与用途
sg	1	1 表示设置算法为 skip-gram
size	100	词向量维数,默认为 100 便于后续计算
window	5	训练窗口大小,一般为 5
min_count	5	字典截断最低频次,默认为 5
sample	1e-3	采样阈值,词频越高越易被采样,默认值 1e-3
hs	0	不使用 HS 方法,采用负采样方法
negative	3	针对负采样 noise words 个数,一般为 3

3.2.2 文档向量表示

LDA2vec 模型的另一部分输入来自于 LDA 模型的输出结果,即主题-词分布矩阵和文档权重。所以在本研究中,同样以预处理后的语料库作为数据集输入 LDA 模型进行训练。Python 中包括 gensim、sklearn 等众多包都有对于 LDA 模型的封装,考虑到后续实验中的困惑度评价指标计算,笔者选择基于 sklearn 来实现 LDA 模型的训练。

主题数目的设置对于 LDA 模型的输出结果有着很大的影响。如果主题数目设置过多,会导致结果不显著;如果主题数目设置过少,则会出现部分词汇对应多个主题的结果。而困惑度(perplexity)通常作为主题

模型的一个主要评价指标，描述了主题划分的确定性如何，能在一定程度上反应模型的优劣。虽然主题数目的选择会影响 perplexity 值的计算，perplexity 值只能作为一个参考，主题数目的确定还需要考虑主观需求。随着主题个数选定的不同，模型的困惑度是不断变化的，具体情况如图 2 所示。综合考量困惑度值和本研究主观需求，将主题数量 K 设置为 15；超参数取默认值进行 LDA 模型的训练以获取主题－词矩阵和文档权重。

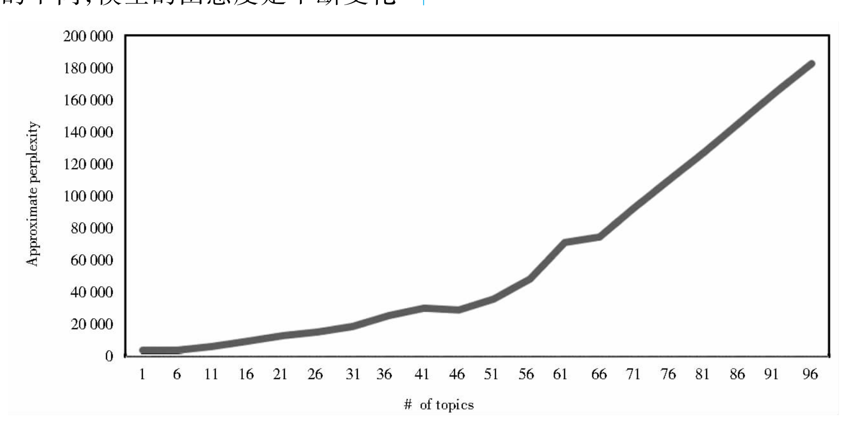


图 2　LDA 模型困惑度值随主题数目变化

3.2.3　基于 LDA2vec 提取主题

LDA2vec 模型的输入包括 3.2.1 中得到的词向量和 3.2.2 中计算得到的文档向量，将其输入 LDA2vec 模型融合训练。

提取到的结果如图 3 所示。按照主题出现概率从高至低排序，共 15 个主题。选取显示每个主题下 top10 概率的主题词，以更清晰、准确的理解每个主题的隐含语义。

```
write p(z): ../output/model/k15.pz
write p(w|z): ../output/model/k15.pw_z

write p(z): ../output/model/k15.pz
write p(w|z): ../output/model/k15.pw_z
procedure time : 4589.50429
================ Topic Display =============
K:15, n(W):4176
p(z)      Top words
0.242885  学习:0.078561 机器:0.043403 研究:0.036719 数据:0.028418 技术:0.027468 方法:0.023497 分析:0.022152 算法:0.021871 领域:0.015923 发展:0.015687
0.148727  模型:0.074640 预测:0.052726 算法:0.038328 学习:0.031373 机器:0.020169 方法:0.015522 分类:0.014320 数据:0.013324 特征:0.013022 回归:0.012882
0.107437  算法:0.034364 特征:0.032720 方法:0.031715 分类:0.025391 学习:0.024914 检测:0.022468 数据:0.021646 机器:0.016591 样本:0.014622 提出:0.013018
0.080647  人工智能:0.053235 发展:0.042506 技术:0.041942 智能:0.025238 系统:0.021223 学习:0.016912 分析:0.016295 领域:0.015744 机器:0.014012 计算:0.011340
0.078303  人工智能:0.035133 发展:0.026566 技术:0.021840 数据:0.019546 研究:0.009859 经济:0.009542 决策:0.009500 风险:0.009376 规划:0.009003 城市:0.008574
0.070740  数据:0.059370 系统:0.027379 分析:0.019251 用户:0.018596 技术:0.014585 预警:0.012679 机器:0.012328 网络:0.012191 平台:0.011978 故障:0.011536
0.051425  诊断:0.031890 模型:0.021795 临床:0.019723 患者:0.019557 预测:0.019391 学习:0.014603 方法:0.013525 分析:0.013235 机器:0.013028 影像:0.012986
0.039415  预测:0.021127 变量:0.020460 算法:0.018004 土壤:0.017390 研究:0.016002 机器:0.014027 模型:0.014000 利用:0.011571 信息:0.011331 学习:0.010957
0.038758  样本:0.029628 对抗:0.026265 分类:0.022468 研究:0.022387 特征:0.019187 方法:0.015146 图像:0.013383 影像:0.013275 学习:0.013031 遥感:0.013031
0.033980  数据:0.031286 课程:0.022694 教学:0.022264 学习:0.020300 学生:0.019502 专业:0.016618 企业:0.016403 创新:0.014960 培养:0.014715 人工智能:0.014408
0.033200  系统:0.025199 测试:0.020871 算法:0.020369 船舶:0.019993 环境:0.018864 机器:0.015791 路径:0.014473 学习:0.013846 跟踪:0.012560 运行:0.009989
0.030564  网络:0.075796 神经:0.037652 识别:0.024881 生成:0.013262 欺诈:0.013093 机器:0.012077 学习:0.010518 结构:0.010349 芯片:0.010010 信号:0.008791
0.024615  预报:0.044148 模式:0.028888 模型:0.016811 学习:0.014206 跌落:0.013213 知识:0.012345 构建:0.012014 效果:0.011600 关系:0.011270 事故:0.011270
0.019302  效应:0.025577 经传:0.022538 预测:0.016974 体质:0.012698 材料:0.012029 dna:0.010380 数据:0.010226 研究:0.009556 基因组:0.009556 层次:0.009195
0.000001  实体:0.000239 语义:0.000239 关系:0.000239 分类:0.000239 研究:0.000239 目的:0.000239 意义:0.000239 信息:0.000239 抽取:0.000239 结构化:0.000239

Process finished with exit code 0
```

图 3　LDA2vec 主题识别结果

3.3　基于 LDA2vec 的主题可视化

利用 pyLDAvis 工具包对主题识别的结果进行可视化展示，能够更加直观地对热点主题结果进行观察和分析，可视化结果见图 4。

基于 pyLDAvis 工具的可视化界面分为了两个部分，页面左侧对识别出的所有主题进行可视化展示，以图形大小代表主题出现概率大小、以图形之间位置关系表明不同主题之间远近关系；页面右侧是对所有主题词概率的可视化展示，浅色条块表示该主题词一共出现的频率，当选中左侧某主题后，会在浅色基础上将某一主题下该主题词出现的频率标深。

基于该可视化图形，能够更加清楚地探析热点主题、突出主题、主题之间关系等内容。从图中结果可以看出，识别出的第一至四主题在所有文献中占比为绝大多数，观察 topic0－3 的特征词，可将其归纳为“算法与方法”“文本分类”“特征检测”与“数据分析”，其中对于机器学习相关算法与方法的研究在机器学习领域中占绝对地位，与其他几个主题也有所关联。第一象限中 8 个主题“医疗应用”“预测分析”“图像”“教学应用”“机械应用”“通信与信号”“预警

系统”“基因应用”与“语义”占比虽然不多，与前 4 个相比差距较大，但彼此之间关联与重合较多，可见机器学习在不同领域的应用是相互可以借鉴、关系非常紧密的。

即从以上两方面的可视化结果来看，基于 LDA2vec 模型提取的主题是内容充实的，主题之间的关系是比较清晰分明的，没有过多的出现重叠、交叉等现象。且利用 pyLDAvis 工具能够很方便地对识别出的主题进行关系、内涵、意义上的深入探索和分析，这对于科研工作者来说是大有裨益的。

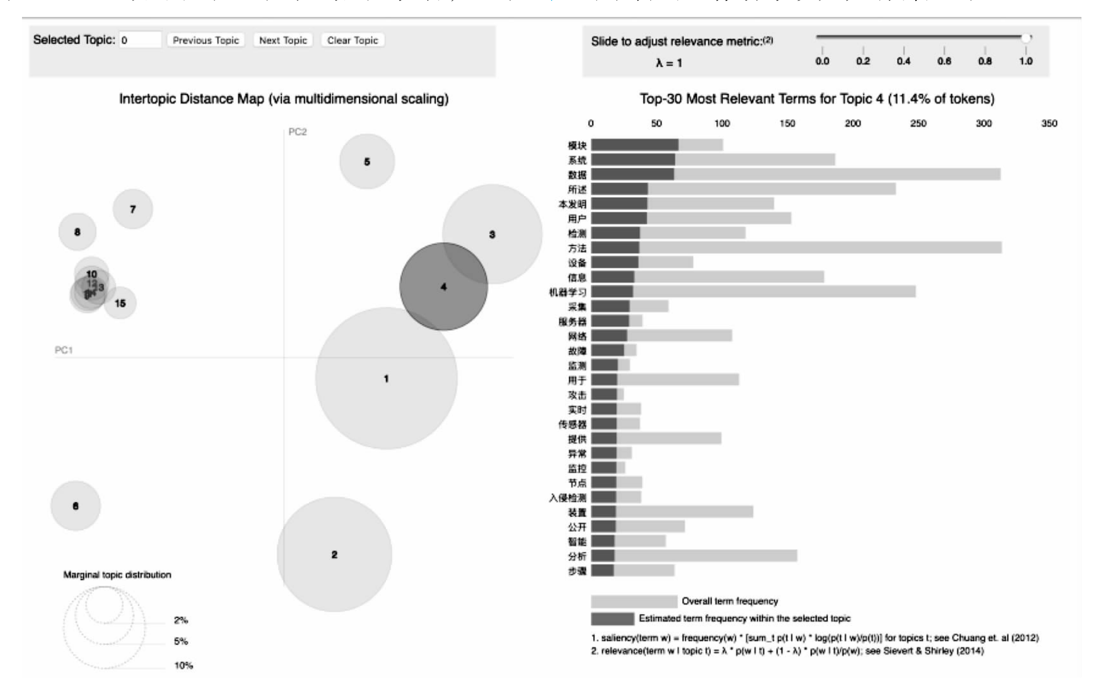


图 4　LDA2vec 主题识别结果可视化

3.4　实验对比评价

C. E. Moody 人在其研究中以 Hacker News 网站的评论数据和一个文本分类聚类经典数据集 Twenty Newsgroups 为实验数据，主要对 LDA2vec 模型进行可行性验证，展示模型识别结果，对其中一小部分结果计算了主题一致性，但并未就模型性能与传统模型进行比较。本次实验将从两个方面对实验结果进行对比与评价分析，验证本文方法的可行性与有效性。一方面基于一个广泛使用的评价指标——困惑度进行评价，困惑度可以理解为对于一篇文章 d，所训练出来的模型对文档 d 属于哪个主题有多不确定，这个不确定程度就是困惑度。困惑度越低，说明聚类的效果越好。另一方面，基于主题一致性（Topic Coherence）指标进行评价，通过对识别到的主题下的特征词之间的相似性关系进行量化评价，这一指标可以反映出识别的主题中哪些是可用的、有价值的。

3.4.1　模型困惑度

在信息论中，困惑度是用来对概率模型预测样本好坏程度进行衡量的一个重要指标。在自然语言处理中，一个语言概率模型可以看成是在整过句子或者文段上的概率分布，其基本思想是给测试集的句子赋予较高概率值的语言模型较好。其公式如下：

$$P(\tilde{W} \mid M) = \prod_{m=1}^{M} p(\overrightarrow{\tilde{w}_m} \mid M)^{-\frac{1}{N}} = \exp -\frac{\sum_{m=1}^{M} \log p(\overrightarrow{\tilde{w}_m} \mid M)}{\sum_{m=1}^{M} N_m} \quad \text{公式(5)}$$

由公式可知，困惑度越小，句子概率越大，语言模型效果越好。

对同样的数据分别进行 LDA 模型和 LDA2vec 模型的训练，基于 Pyhton 利用 sklearn 包中的 lda_perplexity 函数，分别计算两种模型的困惑度值。并将主题数目的 range 设置为[1,100]，间隔为 5，计算并绘制当主题数目从 1 至 100 变化的过程中，两种模型困惑度的变化情况对比图。主题数目取值在 1 至 100 间 LDA 主题模型和本实验采用模型的困惑度值分布曲线，见图 5。

由图 5 可以看出，其中新模型的曲线一定范围内在 LDA 主题模型的下方显示，尤其是主题数目 K < = 40 的情况下。这也是符合大多数情况下进行科研主题识别时的需求，因为对某一学科领域进行科研热点识别，期望达到的目的就是将大量的文档进行有限、有

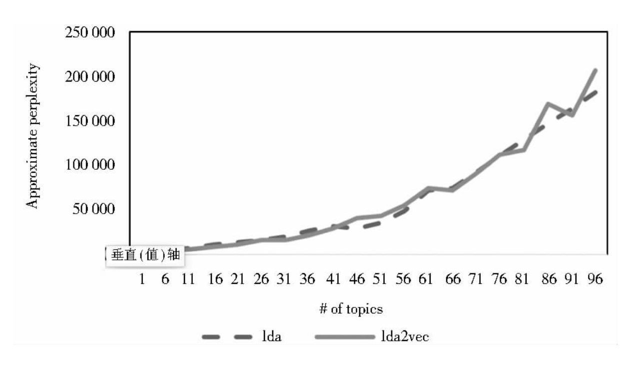


图 5　LDA 与 LDA2vec 模型困惑度比较

效的主题分类，才能更方便地为后续科学研究提供支持与帮助，所以过于多的主题数量也是并不符合我们的需求。由于困惑值越大，模型对样本数据的分类效果越差，反之模型分类效果越好、泛化能力也越强。所以在一定范围内，本实验所采用的模型对于进行科研热点识别来说是更加合适的。

3.4.2　主题一致性

语料库分别在 LDA 与 LDA2vec 算法下进行系列实验，识别出相同数目的热点主题及主题下 top10 的特征词。基于 LDA 模型与 LDA2vec 模型的热点主题识别结果分别如下表 2 和表 3 所示：

表 2　LDA2vec 热点主题识别及归纳结果

主题归纳	主题词 top10
算法与方法	学习 机器 研究 数据 技术 方法 分析 算法 领域 发展
文本分类	模型 预测 算法 学习 机器 方法 分类 数据 特征 回归
特征检测	算法 特征 方法 分类 学习 检测 数据 机器 样本 提出
数据分析	人工智能 发展 技术 智能 系统 学习 分析 领域 机器 计算
金融应用	人工智能 发展 技术 数据 研究 经济 决策 风险 规划 城市
用户分析	数据 系统 分析 用户 技术 预警 机器 网络 平台 故障
医疗应用	诊断 模型 临床 患者 预测 学习 方法 分析 机器 影像
预测分析	预测 变量 算法 土壤 研究 机器 模型 利用 信息 学习
图像	样本 对抗 分类 研究 特征 方法 图像 影像 学习 遥感
教学应用	数据 课程 教学 学习 学生 专业 企业 创新 培养 人工智能
机械应用	系统 测试 算法 船舶 环境 机器 路径 学习 跟踪 运行
通信与信号	网络 神经 识别 生成 欺诈 机器 学习 结构 芯片 信号
预警系统	预报 模式 模型 学习 跌落 知识 构建 效果 关系 事故
基因应用	效应 经传 预测 体质 材料 dna 数据 研究 基因组 层次
语义	实体 语义 关系 分类 研究 目的 意义 信息 抽取 结构化

初步观察来看，上示两个表格的主题词结果，可以较为直观地发现基于 LDA2vec 模型识别的主题词可理解性更高。

3.3 小节中基于 pyLDAvis 的可视化效果的好处是在于可以看出各个主题各自包含的词数以及它们之间距离的远近，使聚类效果更具有可解释性；缺点是无法用数值给出具体好坏。而基于 topic coherence 方法的

表 3　LDA 热点主题识别及归纳结果

主题号	主题词
#0：	模型 预测 风险 学习 机器 基于 数据 患者 临床 方法
#1：	对抗 样本 电网 网页 蜂群 代理 声音 攻击 描述符 含量
#2：	数据 模型 方法 滤波 报告 运动 以及 pi3k 糖尿病 研究
#3：	方法 学习 算法 模型 预测 基于 机器 进行 检测 分类
#4：	学习 机器 研究 数据 深度 算法 进行 应用 方法 基于
#5：	分类 基于 方法 文本 数据 学习 进行 算法 情感 信息
#6：	模型 算法 学习 预测 机器 研究 信息 诊断 方法 航班
#7：	预报 神经网络 模型 异音 效果 模式 质量 芯片 基于
#8：	故障 特征 进行 船舶 网络 android 应用 学习 识别 基于
#9：	算法 预测 设备 学习 机器 数据 聚类 图像 分割 研究
#10：	商品 效应 循经 传导 哈希 web 影响 销量 品牌 机器人
#11：	环境 图书馆 预报 商业银行 客户 识别 城市 划分 操作
#12：	数据 视频 机器 研究 技术 学习 生成 目标 监控 利用
#13：	人工智能 技术 发展 应用 学习 数据 分析 智能 研究 机器
#14：	诊疗 算法 变量 平台 路径 结果 映射 跟踪 船舶 医疗

优点在于用具体数值的方法定量地给出模型的效果好坏化。所以为了进一步验证，笔者利用前文提到的“主题一致性”作为评价指标。由于人们对于主题模型的理解更倾向于属于同一主题的单词在语料库中共同出现的频率，主题一致性度量通过测量主题中高得分词之间的语义相似度来对单个主题进行评分，所以基于该指标的测量有助于区分可解释主题的主题和统计推断的主题。

gensim 0.13.1 版提供了几种不同的计算方法，包括 C_UCI、U-Mass 等，这些计算方式主要的不同在于“共现”的定义不同。

UCI 的计算公式为：

$$score(w_i, w_j, \in) = \log \frac{p(w_i, w_j) + \in}{p(w_i)p(w_j)} \quad 公式(6)$$

它通过在外部语料库（如维基百科中文语料库等）上的滑动窗口中计算单词共现频率来计算单词概率。在某种程度上，该度量可以被认为是对已知语义评估的外部比较。U-Mass 的度量标准定义基于文档共现的分数：

$$score(w_i, w_j, \in) = \log \frac{D(w_i, w_j) + \in}{D(w_j)} \quad 公式(7)$$

其中，D(x,y)计算包含单词 x 和 y 的文档数量，D(x)计算包含 x 的文档数量。U-Mass 指标计算了用于训练主题模型的原始语料库的计数，而不是外部语料库。该指标本质上更具内在性。所以对于我们的评估，笔者决定采用 U-Mass 方法对主题一致性进行测量，且这个评测措施已被证明与人工对主题质量的判断更加匹配。

利用这一评价标准,对划分主题下内容的关联关系进行定量的体现。分别选取两个模型输出的 15 个主题下的特征词汇,利用 U-mass coherence 函数对这些词的主题关联度进行计算,分别得到的结果如表 4 所示:

表 4　LDA2vec 与 LDA 主题一致性结果对比

LDA2vec		LDA	
# of topic	topic coherence	# of topic	topic coherence
0	0.679	0	0.712
1	0.729	1	0.675
2	0.668	2	0.472
3	0.678	3	0.615
4	0.687	4	0.574
5	0.612	5	0.645
6	0.512	6	0.479
7	0.785	7	0.598
8	0.798	8	0.772
9	0.625	9	0.712
10	0.645	10	0.723
11	0.675	11	0.612
12	0.713	12	0.623
13	0.425	13	0.624
14	0.564	14	0.691
average	0.653	average	0.635

由表可见,在大部分识别出的主题范围内,笔者提出的基于 LDA2vec 模型的 topic coherence 值是略高于传统 LDA 模型的,经统计计算得出的平均值 0.653 也略大于 LDA 模型的平均值 0.635。因此,从基于 topic coherence 的定量验证来说,某一主题下主题词的内部关联度也更高,可以更加容易地对其进行理解与归纳,从而得到大致的主题名称,这为科研工作者进行下一步的科研创新工作提供了更高的便利性。

综上所述,本文中采用的基于 LDA2vec 的模型的科研热点识别方法在改善传统识别方法对文本隐含语义表达的缺失基础上,不仅在一定程度上提高了主题识别的精确性,还具备较好的模型泛化能力。

4　总结

4.1　结论

期刊论文、专利文献、政策文本、基金项目数据等,这些内容都是影响科研热点识别分析结果的因素,不同源的文本能够从不同的角度反映关于特定学科领域的研究状态。而在数据时代,能否快速准确的高效获取并理解多源数据,是很多工作的基础与前提。

本文的创新性在于:一是对多源数据融合的应用场景进行了探索。图书情报领域内对于多源数据融合的研究目前较多的是化柏林等[29-31]学者的探索,更多的是从宏观层面上来看图情领域内多源数据融合的意义与方法,但从实际如何应用多源数据来看,缺乏从实际应用场景、应用方案和具体技术实现的细节进行深入的探讨与分析,而本文的研究提供了一种思路。如前所述,不同源的科技文本对于科学研究的贡献度、贡献方向是不同的,有的偏重于理论的研究,有的偏重于方法与技术的实现,有的偏重于领域前沿的探索等。所以本文的创新点之一是将把多源文本进行融合,作为科研热点主题识别的对象,并研究具体实现方法与技术细节。本研究实验中选取了两种不同的数据源——期刊论文数据和专利文献数据,以达到针对某一学科领域进行科研热点主题识别时,能初步将理论与实践相结合。二是对 LDA2Vec 主题模型的应用场景进行了探索验证。目前来说,科研主题热点识别领域内,相对来说更多使用的是传统的 LDA 主题模型,也不乏一些对于 LDA 主题模型的优化与改善,而现有的基于 LDA2Vec 模型的研究更多的是应用于新闻推荐与情感倾向性分析中,本研究创新性的将此模型应用于图情学科中,并提出应用于科研热点主题识别的具体实现方法,某种意义上拓展了模型的实际应用性。

总结来说,笔者提出的基于 LDA2vec 模型的科研热点识别方法,在主题提取上的效果上有相对程度的提升。本研究的模型困惑值在大部分范围内低于传统 LDA 模型的结果,泛化能力更强;且某一主题下主题词的内部关联度也更高,所以可以更加容易对其进行归纳、得到大致的主题名称,为科研工作者的研究提供了更多的便利性。

从整体来说,本文提出的模型与传统 LDA 模型相比,继承了传统 LDA 算法和 Word2vec 词聚类算法优点,对于主题研究具有一定参考价值;且面对多源文本的环境下,该方法也能够有较为不错的表现。通过本文的研究,更好地将 LDA2vec 主题模型方法引入图情学科中来,基于科研热点识别这一应用领域,快速准确地识别蕴含在多源文本中的热点主题,为科研创新提供支撑服务。

4.2　研究局限性

一是在实验数据源选取方面。本文的核心研究内

容是面对多源文本时的主题识别。但是本研究目前只选取了两种数据源——期刊论文数据和专利文献数据进行融合分析,对于其他种类、来源的数据暂未涉及,还未能探讨多种不同类型的数据源对于实验结果的影响。

二是本研究的数据获取和处理方面。由于获取全文过于庞大,本研究中面对的主要是两种数据源——期刊论文和专利文献的题名与摘要内容。也因此,本研究中对于这两种数据的融合借助了科技文献数据库本身对于题名和摘要的标引功能,且这两种文献的功能结构是相似、完整的,即本研究的数据基础是同构的。如果面对的是异质、异构的数据源时,本文前期的数据获取和处理工作需要再进行深层次的探索。

综上所述,科研热点的识别对于科研工作来说意义非凡,笔者进行了一些方法、应用上的探索,但未来还需针对更复杂的多源数据、更高效的识别效果上更进一步研究。

参考文献:

[1] 邱均平,温芳芳. 近五年来图书情报学研究热点与前沿的可视化分析——基于13种高影响力外文源刊的计量研究[J]. 中国图书馆学报,2011,37(2):51－60.
[2] 任红娟. 文献特征融合的的科学结构分析方法研究[J]. 情报杂志,2013,32(7):97－100.
[3] MORRIS S A, YEN G, WU Z, et al. Time line visualization of research fronts [J]. Journal of the Association for Information Science & Technology,2003,54(5):413－422.
[4] SMALL H. Co-citation in the scientific literature: a new measure of the relationship between two documents[J]. Journal of the Association for Information Science & Technology,1973,24 (4):265－269.
[5] 雷晓庆,刘晓雁. 论文关键词特征的统计与分析[J]. 图书情报工作,1998(5):19－20,32.
[6] 曾倩,杨思洛. 国外图书情报学科知识交流的比较研究——以期刊引证分析为视觉[J]. 情报理论与实践,2013,36(10):114－119.
[7] 智库百科[EB/OL]. [2019－02－20]. https://wiki.mbalib.com/wiki/知识单元. 2018－12－02－2019－02－28.
[8] 王晓光. 科学知识网络的形成与演化(I):共词网络方法的提出[J]. 情报学报, 2009,28(4):599－605.
[9] 祝清松,冷伏海. 基于引文内容分析的高被引论文主题识别研究[J]. 中国图书馆学报, 2014,40(1):39－49.
[10] 杨超,朱东华,汪雪锋,等. 专利技术主题分析:基于SAO结构的LDA主题模型方法[J]. 图书情报工作,2017(3):86－96.
[11] 阮光册,夏磊. 基于Doc2Vec的期刊论文热点选题识别[J]. 情报理论与实践,2019, 42(4):110－115.
[12] 赵一方,裴雷,康乐乐. 基于段落信息增益的政策文本主题识别研究[J]. 数字图书馆论坛,2018(11):2－10.
[13] LOWE S A. The beta-binomial mixture model for word frequencies in documents with applications to information retrieval[C]//Proceedings of the sixth European conference on speech communication and technology. Budapest: Dragon System Inc,1999.
[14] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of machine learning research, 2003, 3:993－1022.
[15] 冯佳,张云秋. 基于本体的研究主题语义分析方法研究[J]. 图书情报工作, 2018,62(7):96－103.
[16] 王洪伟,高松,陆颋. 基于LDA和SNA的在线新闻热点识别研究[J]. 情报学报, 2016(10):1022－1037.
[17] 李永忠,蔡佳. 基于LDA的国内电子政务研究主题演化及可视化分析[J]. 现代情报, 2017,37(4):158－164.
[18] 叶春蕾,冷伏海. 基于引文－主题概率模型的科技文献主题识别方法研究[J]. 情报理论与实践,2013,36(9):100－103.
[19] 王连喜. 国内微博研究热点分析及主题挖掘——以计算机和图书情报学科为研究对象[J]. 情报杂志, 2015(4):127－132.
[20] 马红,蔡永明. 共词网络LDA模型的中文文本主题分析:以交通法学文献(2000－2016)为例[J]. 数据分析与知识发现, 2017,32(12):17－26.
[21] 沈思,徐飞,吴鹏. 面向科学研究主题的文献隐含时间信息分析与挖掘[J]. 情报学报,2017,36(4):370－381.
[22] 蒲姗姗. 基于知识互补的科研合作专家推荐模型研究[J]. 情报理论与实践, 2018,41(8):100－105.
[23] 周娜,李秀霞,高丹,等. 基于潜在主题的知识组合分析研究——以传播学为例[J]. 农业图书情报学刊, 2018(9):85－90.
[24] 刘玉文,吴宣够,郭强. 网络热点新闻焦点识别与演化跟踪[J]. 小型微型计算机系统, 2017(4):738－743.
[25] 张聪,易秀双,朱明浩,等. 一种基于Spark学术研究热点的挖掘方法[J/OL]. 计算机工程,2019. [2019－02－20]. http://kns.cnki.net/kcms/detail/31.1289.TP.20190129.1332.005.html.
[26] 关鹏,王曰芬. 学科领域生命周期中作者研究兴趣演化分析[J]. 图书情报工作, 2016,60(10):116－124.
[27] HOFMANN T. Probabilistic latent semantic analysis[C]//Fifteenth conference on uncertainty in artificial intelligence. Berkeley: Morgan kaufmann publishers Inc. ,1999:289－296.
[28] MOODY C E. Mixing dirichlet topic models and word embedding to make LDA2vec[EB/OL]. [2019－02－20]. http://arxiv.org/abs/1605.02019.
[29] 化柏林,李广建. 大数据环境下的多源融合型竞争情报研究[J]. 情报理论与实践, 2015,38(4):1－5.
[30] 化柏林,李广建. 大数据环境下多源信息融合的理论与应用探

讨[J].图书情报工作,2015,59(16):5-10.
[31] 化柏林.多源信息融合方法研究[J].情报理论与实践,2013,36(11):16-19.

作者贡献说明:

裘惠麟:设计论文整体结构,撰写论文;

邵波:提出论文研究思路与框架,指导论文修改定稿。

Research on Identification Methods of Scientific Research Hotspots Under Multi-source Data

Qiu Huilin[1] Shao Bo[1,2]
[1] School of Information Management, Nanjing University, Nanjing 210046
[2] Nanjing University Library, Nanjing 210046

Abstract: [**Purpose/significance**] In scientific research, identifying mining scientific research hotspots from different sources of scientific literature is of guiding significance for carrying out the next scientific research work. It aims to quickly and accurately identify hot topics contained in multi-source texts through the model method proposed in this study, and provide support services for scientific research innovation. [**Method/process**] This paper proposed a method based on LDA2vec model for multi-source text research hotspot identification and built a model for scientific research hotspot identification. This method combined the advantages of LDA topic model on implicit semantic mining and the context of Word2Vec word vector model. Taking the scientific literature in the field of machine learning as an example, the model extraction degree (perplexity) and topic coherence (topic coherence) were used to compare the topic extraction effects of LDA2vec and LDA in the context of multi-source text. [**Result/conclusion**] After experiments, the results show that the method proposed in this paper is feasible and can be improved to some extent in the face of multi-source data. The method can relatively quickly and accurately identify the hot content in the multi-data source text, make up for the shortcoming of the single analysis data source for subject detection, and enrich the practical application of the multi-data source fusion theory system.

Keywords: topic model LDA2vec research hotspot LDA word2vec multisource data fusion